An XGBoost-SHAP Framework for Interpretable Travel Time Prediction at MARTA

Kavin Somu

Master of Science in Analytics

Georgia Institute of Technology

ISYE 6748 Practicum

07/17/2025

Problem Statement:

Metropolitan Atlanta Rapid Transit Authority (MARTA) runs over 113 bus routes throughout the city. On Time Performance (OTP), the percentage of bus timepoints operated on or within the schedule, is one important measure of service quality. A timepoint is a designated stop along a bus route where the bus is scheduled to depart at a specific time. All timepoints are bus stops but not all bus stops are timepoints (See A1 for additional information on timepoints). The OTP target for MARTA is to have more than 78.5% on-time performance. Improving OTP is a complex, system-wide problem influenced by a vast array of variables, which can be classified as:

- **Exogenous factors:** traffic flow, weather, passenger activity
- **Endogenous factors:** operating behavior, mechanical condition of buses

As the factors are interrelated, an important initial step to minimize OTP is to focus on estimating bus *travel times* between timepoints. Travel time is a primary factor for OTP as a bus is considered "on time" at a given timepoint crossing if it departs that timepoint between 30 seconds before and 5.5 minutes after the scheduled departure time. The travel time over which the bus traverses a given timepoint segment (route covered between timepoints) is a crucial factor impacting whether or not the bus will be "on time" at each subsequent timepoint. Moreover, travel times offer independent observations, which are better aligned with analytical modeling assumptions and have a direct impact on buses being OTP compliant. The primary aim of this project is to identify influential factors impacting MARTA's travel times (likely extending to OTP) by:

1. Developing a prediction model to predict travel time between timepoint segments along 7 MARTA's routes with among the lowest OTP.
2. Profiling the relative significance of influencing factors on bus travel times.

These insights will inform how MARTA prioritizes the allocation of resources to enhance on-time performance by identifying areas with greatest opportunities for improvement.

Literature Review:

Utility and History of Bus Travel Time Prediction

Travel time prediction in the public transit sector has been of research interest for many years. Various methods have been developed ranging from naive historical averages to sophisticated machine learning methods. The next section provides a concise overview of the key studies and methods of relevance to predicting bus travel times.

Traditional methods for predicting travel time have laid the foundation public transit analytics; these models include historical average models, linear regression, and Kalman filters (Chiang et al, 2020). These models typically rely on a limited set of features: speed, distance, dwell time and historical patterns. Models using these methods perform well in stable traffic environments, but under dynamic urban environments, their performance declines.

Chiang goes on to explore other machine learning techniques in his paper such as artificial neural networks, support vector machines, and deep learning. These models improve upon the traditional methods, especially when GPS and timestamp data are provided. However, a majority of the reviewed studies in this work heavily rely on exogenous features such as traffic conditions, weather or road network. Since these factors are outside of a transit agency's control, they offer limited guidance on what actionable steps can be taken to address them.

Despite a growth in machine learning models in the industry, there remains a gap in endogenous features, which transit agencies can act upon. Our work aims to address this gap by profiling the relative importance of all factors that contribute to buses' travel time. By incorporating endogenous factors, MARTA is able to direct resources, coaching, and scheduling changes where they can improve on-time performance.

XGBoost Accurately Predicts Bust Travel Times Compared to Other ML-Based Models

In 2022, a study (Zhu et al.) employed an XGBoost model to predict bus travel time from between stations by leveraging 28 days of bus data in Guangzhou. They found evidence that XGBoost performs better than alternative machine learning models including KNN, BP Neural Networks, and LightGBM because its MAE and RMSE values were lower. Additionally, they also prioritized model travel times at the segment level rather than the entire route allowing the model to learn more localized relationships between features and travel time. This study is extremely applicable in that it has established the effectiveness of XGBoost compared to other traditional and machine-learning based approaches in travel time prediction.

Commented [KM1]: According to what criteria? What is "better" here?

Ashwini et al. (2022) demonstrated the ability of XGBoost for bus travel time prediction in sparse data settings. Their work showed that XGBoost outperformed traditional linear models by capturing nuanced, nonlinear trends and leveraging temporal features such as time of day and historic trip information. They developed two XGBoost models, which had R^2 values of .8 and .71. These results substantiate XGBoost as a high accuracy model in public transit settings when paired with thoughtful feature engineering. These results informed the current project's decision to use XGBoost as the modeling technique for MARTA bus travel time prediction.

While these studies have made a strong case for using XGBoost for modeling bus travel time, these studies were conducted in China and India, where infrastructure and traffic patterns differ from those in the United States. This project addresses a gap by applying XGBoost to a city in the United States, which can serve as a basis for future domestic studies. In addition, we address another gap by focusing on the lowest performing routes. We hypothesize that these routes exhibit less predictable conditions, making them more valuable for predictive modeling. By addressing more challenging route, we build a model that is flexible and robust under varying conditions.

Hyperparameter Tuning in XGBoost Models

Anggoro and Mukti (2021) conducted a study that compared the Grid Search and Random Search methods of XGBoost hyperparameter optimization for predicting chronic renal failure. The study applied various preprocessing techniques such as normalization and oversampling to improve model performance on the data. Across these scenarios, grid search consistently produced higher accuracy and F-measure scores than Random Search. The present project was motivated by that to employ Grid Search as a model optimization method.

Commented [KM2]: Why is grid search more accurate?

Default parameter values for the grid search for this project were established by following guidelines from recent literature on XGBoost hyperparameter ranges (Bentejac et al., 2020). This method guaranteed that the calibration procedure adhered to established best practices and was based on empirically validated parameter ranges.

Model Interpretability using SHAP

Zhang et. Al (2023) combined SHAP with XGBoost to explain factors that affect landslide susceptibility. They examined environmental variables such as elevantion, rainfall, slope, and lithology and implanted SHAP to quantify the contribution of these variables on the classifying landslide susceptibility. For instance, they discovered that lithology and mountainous regions were significant contributors, but lithology and flatter regions had a lesser impact. SHAP provided meaningful explanatory value for XGBoost models, especially when feature importance varies across different contexts.

Given the pairing of SHAP and XGBoost has been demonstrated to provide value in other areas of research, bus travel time modeling would benefit from their synergies. SHAP will provide insight into whether exogenous or endogenous factors have a greater impact on the travel time. MARTA can then target which factors to prioritize in implementing policy change.

Data Sources/Feature Engineering:

Factor Type	Factor	Feature	Data Source	Description/Calculation Details
Endogenous	Operator Behavior	Coaching Sessions	OTP PBI Report	Number of coaching sessions completed by an operator as of April 18, 2025.
		On Time Rate	OTP PBI Report	Used rolling window to calculate number of trips on time divided by total number of trips between all dates an operator worked and the date 6 months prior to those.
		Days Since Last Coaching	OTP PBI Report	Difference in days between date of operation and day of last coaching session.
		Work Experience	TMDM Operator	Difference in days between the operation date and the activation date of an operator.
	Bus Maintenance	Total Symptoms	Incidents PBI Report	Used rolling window to calculate number of recorded symptoms between all dates a bus traveled and the date 6 months prior to those.
		Severe Symptoms	Incidents PBI Report	Used rolling window to calculate number of recorded "High Severity" symptoms between all dates a bus traveled and the date 6 months prior to those.
		Service Length	TMDM Vehicle	Length of time the vehicle has been in service. The difference in days between the operation date and the first date in service of a bus.
Exogenous	Passenger Flow	Made Stop Ratio	TMDM via Automatic Passenger Counters	The number of stops recorded along a timepoint segment divided by the total number of stops along that timepoint segment.
		Passenger Activity	TMDM via Automatic Passenger Counters	The sum of passengers that boarded and alighted the bus along a timepoint segment
	Weather	Temperature	Meteostat.net	Temperature in Fahrenheit.
		Precipitation	Meteostat.net	Precipitation in inches.
	Traffic	Traffic Travel Time	RITIS	The sum of traffic travel time amongst all RITIS segments corresponding to a timepoint segment.
		Route Pattern	TMDM Adherence	Concatenated the route, direction, and pattern fields.
		Timepoint Segment (Model 3)	TMDM Adherence	Identification number of a segment on a particular bus route

Commented [KM3]: Switch

Table 1: Summary of Features

This project utilizes a set of data sources for modeling bus travel time on MARTA's Route×Direction×Pattern combinations with among the lowest historical OTP. For the purposes of OTP calculations, MARTA only includes trips that fall in the window of 10 minutes early and 20 minutes late. Currently, there isn't a good way of flagging data errors, and observations that fall outside this window are more likely to be data errors than true schedule deviations. Therefore, the dataset will be filtered on trips that adhere to this criteria.

Several features from these sources were engineered to serve as inputs for the model as shown in Table 1. The operations were designed to produce informative variables that indeed capture the static and temporal characteristics that are relevant to determining MARTA bus travel times.

TMDM and Route Pattern Identification

MARTA's Transit Master Data Mart (TMDM) is the database from which travel times between timepoint segments are extracted for each trip corresponding with the Route×Direction×Pattern combinations of interest. Additionally, TMDM offered essential trip-level attributes, including route, direction, pattern, operator ID, vehicle ID, and stop-level segment information. These fields served as the foundation for joining external datasets and creating the timepoint-segment-level observations used within the predictive model. I designed a distinct *route pattern* identifier by concatenating the three columns: route, direction, and pattern. The route pattern serves two crucial functions: it acts as an input feature in the prediction model, and as a shared key for the integration of TMDM data with external traffic data in the RITIS. This step ensures that attribute data of buses is in accordance with the corresponding traffic data.

RITIS Traffic Conditions

External traffic conditions impacting bus segments are collected via traffic speed data from the Regional Integrated Transportation Information System (RITIS). Since RITIS reports traffic at fixed roadway segments that don't correspond one to one with timepoint segments, a route-matching process was needed to connect RITIS traffic with MARTA's routes in TMDM. The MARTA transit analytics team previously reconstructed bus routes as RITIS segment sets to capture external traffic conditions. In cases where RITIS segments were greater than TMDM segments, scaling factors were applied to adjust for such differences. For instance, a scaling factor of 5/6 was used to proportionately modify the RITIS travel time if one segment of RITIS was 2.4 kilometers and the corresponding TMDM segment was 2.0 kilometers. This modification holds on the assumption that traffic conditions within the supplementary or missing segments of the RITIS segment are the same as those in the timepoint segment. To estimate these scaling factors, the exact distances were measured by mapping the RITIS and TMDM segments via Google Maps.

Each timepoint segment's corresponding RITIS segments were added up to estimate the total observed RITIS distance. It was then possible to calculate an effective scaling factor that was subsequently utilized to predict the respective RITIS travel times. Please refer to A3 for a detailed overview of this process. In the attempt to create a traffic-feature that closely approximates real on-road traffic conditions, the updated travel times for each RITIS segment were aggregated at route pattern, timepoint segment, and 15-minute interval levels. This *traffic travel time* attribute enables the model to capture the variation in travel times resulting from external congestion along a segment.

Weather Data from Meteostat

Weather data were extracted using the Meteostat Python library from meteostat.net, which summarized *temperature* and *precipitation* data from the nearest four metro-Atlanta stations. The data are employed to approximate the weather impact on travel time but may fail to observe microclimate impacts along specific routes.

Passenger Activity Features

The TMDM database also reports automatic on-board passenger counter data, including boardings and alightings which may have an impact upon dwell time, which was used to collect passenger activity at each stop and aggregated to each timepoint segment. Two measures of passenger flow are developed to quantify passenger activity. These are used to record passenger activity's effect on travel time variance. The *made stop ratio* is calculated by using the number of stops actually made divided by the number of planned stops within a timepoint interval. Boardings and alightings at each stop were also added up to create *passenger activity,* one of the measures for total passenger movement. This variable follows the cumulative movement of passengers at each timepoint. For both variables, we relied on the assumption that travel time would not be affected by anything at the final stop since the trip is marked as complete upon arrival. The final stop of all trips was not included in this calculation to make sure that only activities pertinent to travel time were taken into consideration.

Operator Behavior and Performance

The characteristics of operators driving the vehicles on a given trip, including their history of on-time performance on suitable routes, time as operators, and history of operator coaching were pulled from existing Power BI reports at MARTA and are represented using operator activity data. To quantify the behavior of operators, the *on time rate* feature was developed based on past performance of all operators. The rolling window approach was used to calculate the ratios of on-time trips for every trip in the dataset for the previous six months. The approach creates performance metrics that change over time and include the latest six-month history of each operator on the day of operation. The analysis prioritized trips operated on routes with an on-time performance (OTP) of greater than 78.5% to guarantee that the proportions account for substantial service quality. If qualifying trips for a particular operator cannot be determined using this high-performing criterion, the analysis defaulted to the inclusion of trips operated on routes with an OTP between 70% and 78.5%. Finally, *work experience* was the period in days between the operators' activation date and trip date. This variable is the possible influence of operator experience and tenure on travel time variability.

Vehicle Health and Maintenance History

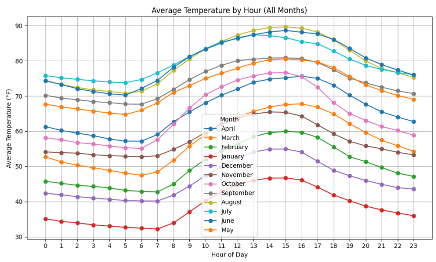
Vehicle service incident records, which encompass the history of each vehicle's operating journeys, including the duration of its service and any mechanical issues it has encountered, were also extracted from existing Power BI reports at MARTA. The dynamic mechanical condition of the vehicle for each operating date was also documented by converting the history of bus maintenance into temporal form. The *total symptoms* and the *severe symptoms* over the past nine months were computed for each operating date. The counts are only updated with the arrival of new maintenance incidents. It is necessary to include maintenance attributes because buses with frequent or severe mechanical failure will more likely be delayed, resulting in higher travel times. Furthermore, *length of service*, the number of days the bus has been on the road since its first date of service, is a feature that indicates the impact of vehicle aging on its travel time.

Feature Characterization:

This section will give an overview on a specific subset of features utilized in the model. These variables have been selected to cover a wide range of features utilized in the model. Additional visualizations on the complete set of features will be presented from A4 to A8 in the Appendix for reference.

Weather Conditions

Temperature was added as a feature since it can potentially affect travel time. It exhibits a parabolic daily profile (Figure 1), and the minimums occur in the early morning and late evening hours and the maximums occur in the midday hours. Increased passenger use may be correlated with increased midday temperatures, which will result in increased trip times. The temperature distribution is left skewed and peaked at around 70–75°F (Figure 2). The temperature range is wider during cold months. The demand for passengers may be higher during the summer months, which may result in longer travel periods for buses due to the increased passenger activity.



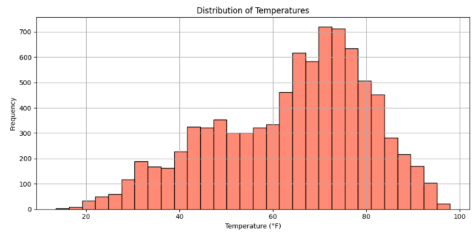


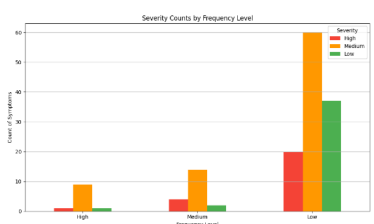
Figure 1: Chart showing parabolic pattern of temperatures by hour.

Figure 2: Chart showing temperature distribution in Atlanta for 1 year.

Bus Maintenance

Bus maintenance issues were defined between August 2023 and April 2025 to assess their potential impact on travel time. Symptoms were categorized into two criteria: frequency, which is referred to as number of events, and severity, which is referred to as average delay per symptom. The severity categories were high (more than 90 minutes delay), medium (60 to 90 minutes delay), and low (less than 60 minutes delay). Similarly, the frequency categories were categorized as high (more than 200 occurrences), medium (50 to 200 occurrences), and low (less than 50 occurrences).

Most high-frequency symptoms, as shown in Figure 3, fall within the medium severity category. This suggests that high frequencies of mechanical issues can still be a major contributor to travel time delays. The *length of service* was experimented with as a potential variable that could be affecting performance. The length of service distribution reveals that many of the buses have been in service for a long time. This can increase travel time as the older buses would be more susceptible to delays caused by mechanical failures.



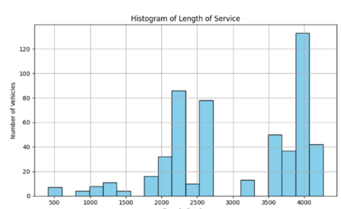


Figure 3: Bus Maintenance Symptom Counts by Frequency and Severity

Figure 4: Distribution of Bus Service Age

Operator Performance

Operator *on time rate* was assessed by using a rolling six-month window to estimate the percentage of timepoints made on time for each operator-date pair. The methodology offers a dynamic measure and it is updated every time new trips are completed, and therefore, it captures both recent and historical performance. Looking at Figure 5, on-time performance is distributed where the majority of the observations fall between 0.85 to 0.95 proportion on time range. For poorer-performing operator-days, there is a left-skewed distribution, meaning a broad range of values. Depending on whether the operators were early or late, on-time performance would have a positive or negative impact on travel times.

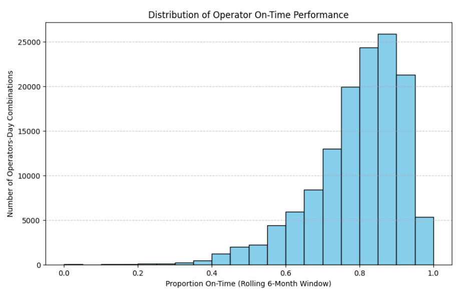


Figure 5: Distribution of Operator On-Time Performance (Rolling Six-Month Window)

Passenger Activity

Passenger flow information is recorded in two variables: *passenger activity* and *made stop ratio*. The *made stop ratio* records the proportion of scheduled stops along a timepoint segment that the bus made. This metric captures the variability of stops, in that the travel time may be reduced if no stops are made, but it may be longer if the stopping rate increases. *Made stop ratio* distribution across timepoint segment trips is right-skewed, as demonstrated in Figure 6. The majority of trips have low ratios, but the set of trips disperses with higher ratios. The boardings and alightings at every timepoint segment were summed up to build *passenger activity*. Higher passenger activity should lead to longer dwell times, which should further increase travel time. The distribution of *passenger activity* is skewed to the right, and there are lower activity levels for the majority of timepoint segment trips.

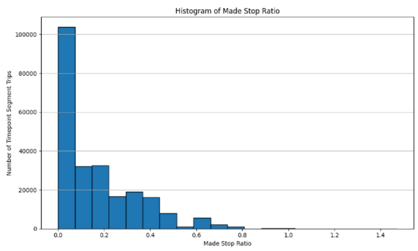


Figure 6: Distribution of Made Stop Ratio Across Timepoint Segment Trips

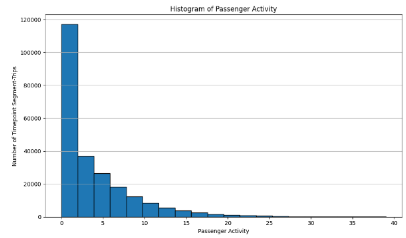


Figure 7: Distribution of Passenger Activity Across Timepoint Segment Trips

External Traffic Conditions

To identify periods of congestion most likely to affect bus travel times, *traffic travel time* was analyzed for all routes by time of day. The actual impact on travel time is highly route and timepoint-specific, though the average traffic travel time has definite peaks, around 8:00 AM and again during 4:00 PM to 6:00 PM. The chart illustrates the fluctuation of different segments using a shaded standard deviation band. Whereas the underlying data indicates traffic conditions at every 15-minute interval, Figure 8 is averaged to hour-long means for illustration purposes. With more congestion and variability in travel times during peak hours, bus travel times would be longer and more variable.

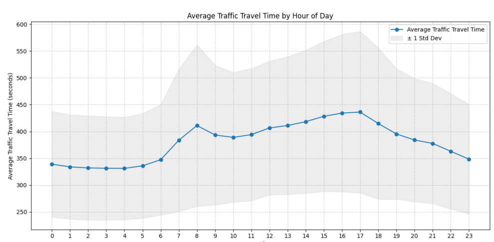


Figure 8: Average Traffic Travel Time by Hour of Day

Travel Time with Dwell Distribution

Figure 9 shows the distribution of *actual travel time,* which includes the travel time between along a timepoint segment, including time spent when the operator is not driving the bus. The distribution is roughly normal with some right skewness, meaning the variability amongst longer trips is wider. The skew may be caused by several of the factors that will be examined in the model. The average travel time centers around 600 seconds (10 minutes), providing an idea of where the predicted values are expected to lie. It is crucial to comprehend this distribution, as it offers context for the model's prediction target, emphasizing the variability in travel durations that the model must learn to capture.

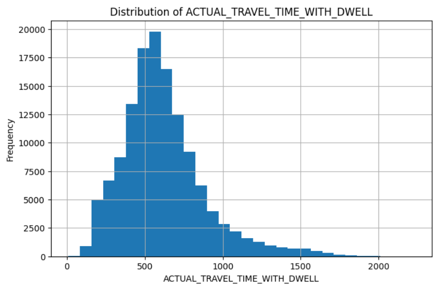


Figure 9: Distribution of Travel Time with Dwell

Methodology:

The MARTA bus travel time predictive model was built in this study via a step-wise procedure. An initial selection of the seven low OTP bus routes was made through a route selection process that concentrated the analysis on the most critical service areas. We then combined internal MARTA data with external data, thereby building an extensive dataset for modeling. XGBoost is a machine learning algorithm that was chosen for its ability to identify nonlinear relationships within enormous datasets in predictive modeling. The parameters of the model went through a hyperparameter tuning phase to determine the best parameters to render the model as accurate as possible. Lastly, the model output was interpreted and relative contribution of every factor towards bus travel times was evaluated using SHAP (SHapley Additive exPlanations) values.

Commented [KS4]: Add subsections to Methodology and other sections.

Route Selection

Routes were evaluated at the route ID, direction, and pattern level for potential candidates to be modeled. Three metrics were used to ascertain suitability:

1. RITIS overlap: The measure of how much of each timepoint segment's planned distance (TMDM) is represented by RITIS-mapped segment lengths, summed across segments to the route level (A2).
2. TMDM data completeness: measures how consistently the data for all timepoint segments along the entire route are available for a given trip for a given route.
3. OTP: mean on time performance for a route.

Commented [KS5]: Ger some insight on these calculations and add a little more so its clear.

As the distance captured by both TMDM and the RITIS datasets can vary for different trips, the mean TMDM to RITIS distance ratio was calculated to assess the overlap between both data sources. With the variation in recorded distances, we also calculated the standard deviation of the mean ratio calculations to reflect the TMDM data completeness. If a trip includes all expected timepoint segments and summed distances match across trips, the standard deviation will be near zero, indicating high data completeness. On the other hand, if trips are missing segments in RITIS, the standard deviation will increase, meaning standard deviation reflects low data completeness.

The metrics were normalized using a Min-Max scaler and then rescaled to correct for preserving its relative spread. Mean ratio, average OTP, and standard deviation of TMDM completeness were weighted at 45%, 35%, and 20% respectively to capture the modeling priorities. A composite score was computed by adding weighted adjusted values; lower scores provided a higher chance of inclusion. To provide sufficient data for modeling, routes with low service frequencies were dropped from consideration in the final route choice process.

Commented [KS6]: Include the pictures from powerpoint

	ROUTE_ABBR	ROUTE_DIRECTION_NAME	PATTERN_ABBR	Mean TMDM to Ritis ratio	Ratio Std	AVG(OTP)	Composite_WeightedMeanAdj
302	832	Westbound	WB	1.030204	0.095620	0.606500	0.293282
117	78	Westbound	IB	0.935931	0.055107	0.588871	0.299013
237	181	Eastbound	IB-OAK	1.020509	0.109326	0.645794	0.343342
61	36	Westbound	WB	1.038296	0.045659	0.680382	0.356121
236	181	Eastbound	IB	0.964469	0.132075	0.623223	0.375713
279	197	Westbound	WB	1.003684	0.060352	0.751729	0.393234
202	140	Southbound	SB	0.992919	0.031137	0.780264	0.405747

Figure 10: 7 Routes Selected for Data

Data Consolidation

The final modeling dataset was acquired using both external and internal data sources surrounding the primary TMDM as shown in A9. Trip IDs, route patterns, timepoint segments, and calendar dates were some selected keys used to join all datasets to TMDM. This process ensured that the objective variable (travel time) and the entire set of explanatory variables were included in every observation of a timepoint segment in the final dataset.

XGBoost

For its proven strengths in terms of computational efficiency and predictive accuracy, XGBoost was selected as the predictive modeling method for the project. According to Zhu et al. (2022), XGBoost is based on the classical Gradient Boosted Decision Trees (GBDT) framework. Support for multiple base learners (e.g., trees, linear models, and dropout-based approaches), incorporating a regularization term in the objective function to limit overfitting, and utilizing a second-order Taylor series expansion to improve optimization precision are some of the principal improvements XGBoost

Commented [KM7]: Just make sure figures are large enough to be legible.

provides. With feature splitting and residual iterative fitting, this blend allows XGBoost to transform weak classifiers into a strong ensemble model in many cases. Figure 11 provides a representation of our XGBoost model. The model feeds the input features into the XGBoost model that learns patterns from the data, and then the model generates predicted travel time for each timepoint segment.

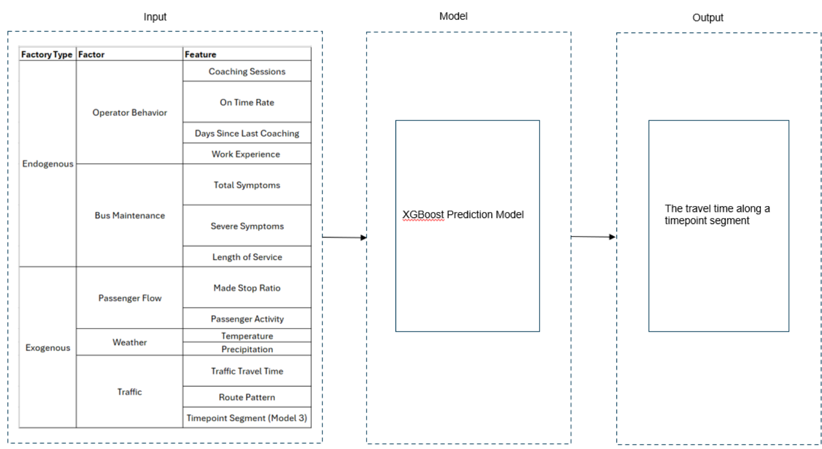


Figure 11: XGBoost Based Prediction Model

Hyperparameter Tuning

A systematic hyperparameter tuning procedure was undertaken to tune the performance of the XGBoost model. Because of its utility in enhancing predictive accuracy, Grid Search with Cross-Validation was used as the optimizing technique. To establish baseline values and candidate ranges, literature-based recommendations that already existed were used to support an exhaustive and fair search. The final choice of hyperparameters using this method is presented in Table 2.

Parameters	Values
n_estimators	700
learning_rate	0.05
max_depth	7
min_child_weight	19
subsample	1
colsample_bytree	0.707
gamma	0
reg_alpha	1
reg_lambda	1

Table 2: Hyperparameter Values Selected via GridSearch

SHAP

We utilized SHAP, a common explainable machine learning approach, to support the interpretability and transparency of our XGBoost model. By measuring each feature's contribution toward each individual prediction and the entire model's behavior, SHAP offers both local and global insights. One of the best aspects of SHAP is that it is based on cooperative game theory and regards every feature as a "player" in a game that has some effect on the model's prediction. By allocating each feature's contribution in a balanced and fair manner and considering all possible combinations of features, SHAP ensures consistency and interpretability. Using visualizations like waterfall plots, bar plots, and beeswarm plots, SHAP enables us to analyze predictions at a individual and aggregate scale.

Evaluation and Results:

In order to measure the performance of the model, the dataset was divided into three sets: training (70%), validation (20%), and test (10%) sets. The XGBoost model was trained on the training set and cross-validation via GridSearchCV from sklearn package was utilized for the hyperparameter tuning. The validation set was utilized to detect possible overfitting and plot the training and validation sets' loss curves. In addition, SHAP analysis was performed on the validation set to perform model diagnostics and identify model behavior trends. The test set, which was not used in the previous steps, was used to evaluate the accuracy of the model. The model's performance was measured using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and R^2 to assess how well the model performed on the test set.

Commented [KM8]: Name the package here

Commented [KS9R8]: from sklearn.model_selection import GridSearchCV

Learning Curve Results

The XGBoost model's RMSE curve is shown in Figure 12 for training and validation over boosting cycles. The RMSE for the training and validation models significantly reduces in the early stages, indicating that the model is learning rapidly. The gap between the two curves widens due to the validation RMSE leveling off at approximately 50 boosting cycles as training RMSE continues to drop. Such divergence shows that overfitting has occurred because the model keeps on optimizing on the training set without making any corresponding optimizations on unseen data. The model utilized early stopping with a patience of 50 rounds to cut the boosting rounds to avoid overfitting because there was no additional validation improvement noticed.

Commented [KM10]: Early stopping parameters?

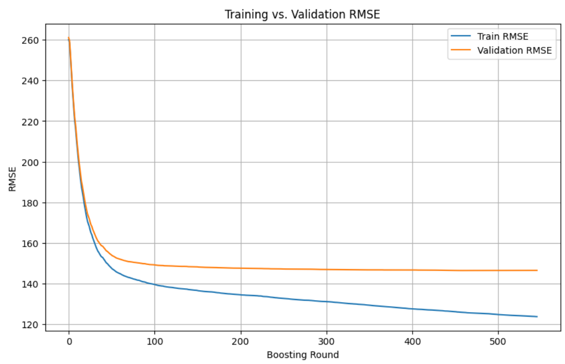


Figure 12: Training vs Validation RMSE Learning Curves for Selected Model

Metric	Accuracy	Meaning
MAE	103.34	On average, predicted travel times deviate from actual values by 1 minute and 43 seconds.
RMSE	166.60	Average error increases to 2 minutes and 47 seconds due to larger predicted values in the model.
MAPE	17.07%	Predicted travel times are off by 17.1% on average from the actual travel time.
R^2	0.6328	The model explains 63.3% of the variation in travel time.

Commented [KM11]: Awesome table.

Table 3: Evaluation Metrics for Original Model 1

Accuracy Metrics for Models 1 & 2

We evaluate the accuracy measures on the test data set for model 1 in Table 3. The model's MAE of 103.34 seconds indicates that the observed actual values and predicted bus trip times were, on average, within 1 minute and 43 seconds. The RMSE captures a slightly greater average error of approximately 2 minutes and 47 seconds, meaning that the predictions have larger deviations. Since the RMSE penalizes these larger mistakes that could be caused by unforeseen delays, it was important to examine both MAE and RMSE. These less frequent but significant deviations must be considered in transportation modeling, where variability can be high.

MAPE of the model is 17.07%. This precision shows that there is still variability present even though predictions tend to be precise in depicting real outcomes. This is most likely due to unpredictable urban traffic conditions. The model's R^2 value of 0.6328 shows that the variables account for about 63.3% of the variation in bus travel time. Even though the model demonstrates a reasonable fit, these results highlight the complexity of predicting bus travel time. Together, these measures yield a comprehensive assessment of the model's performance.

Metric	Accuracy	Meaning
MAE	102.67	On average, predicted travel times deviate from actual values by 1 minute and 43 seconds.
RMSE	165.48	Average error increases to 2 minutes and 45 seconds due to larger predicted values in the model.
MAPE	16.91%	Predicted travel times are off by 26.9% on average from the actual travel time.
R^2	0.6333	The model explains 63.3% of the variation in travel time.

Table 4: Evaluation Metrics for Model 2 with log transformations and interaction variables

Given the skewness and dependencies of some the variables, we implemented log transformation and interaction features for model 2. *Passenger activity* had a skewness value 1.61, which led to using a log transform to reduce the effect of extreme values. Additionally, several interaction features were created based on domain knowledge: (*Made Stop Ratio x Passenger Activity*), (*Route Pattern x Traffic Travel Time*), (*Route Pattern x Made Stop Ratio*), (Route Pattern x Passenger Activity). Interaction variables were created based on the assumption that route dynamics may amplify or mitigate the effects on travel time. Table 4 displays the model results after adding these new features. While the accuracy metrics showed some improvement, the gains were not substantial. Therefore, we used model 1 for SHAP analysis to diagnose any issues with the variables

Model 3 and SHAP Diagnostics

Since the additional transformed features didn't improve performance, we turned to SHAP for model diagnostics to identify issues with Model 1. We performed these diagnostics on the validation set because we wanted to separate this from evaluating its model accuracy on the test set. Figure 13 shows a beeswarm plot for model 1. The color gradient indicates the magnitude of the features value (blue = low, red = high). For example, *traffic travel time* shows a pattern where low values reduce travel time while high values increase travel time (SHAP values will be explained in more detail in the next section). However, *made stop ratio* revealed an unexpected pattern. Many red points appeared on the left side of the plot, suggesting high stop ratios lowered predicted travel time, which contradicts expectations. It turns out that there were time point segments with shorter travel times but had very high made stop ratios. The model learned an incorrect pattern that when buses make many stops their travel time is shorter. This revealed that *route pattern* was limited in capturing the variability amongst different segments. We next shifted

Commented [KM12]: How skewed was the travel time variable and did you try log transforming that?

Commented [KS13R12]: Skewness: 1.1873

Commented [KM14]: Did you try using time of day and day of week as a feature?

to building model 3 where we substituted *route pattern* with *timepoint segment* to provide the model with greater granularity.

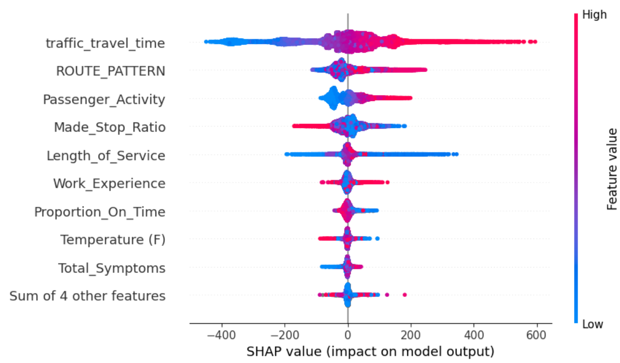


Figure 13: Beeswarm Plot for Model Diagnostics on Model 1

Metric	Accuracy	Meaning
MAE	88.38	On average, predicted travel times deviate from actual values by 1 minute and 28 seconds.
RMSE	144.03	Average error increases to 2 minutes and 24 seconds due to larger predicted values in the model.
MAPE	14.71%	Predicted travel times are off by 14.7% on average from the actual travel time.
R^2	0.7222	The model explains 72.2% of the variation in travel time.

Table 4: Evaluation Metrics for Model 3 with Timepoint segment as an explanatory variable

Replacing *route pattern* with *timepoint segment* in Model 3 led to a significant improvement in model performance. The MAE decreased to 88.38 seconds, meaning the average prediction was about 1 minute and 28 seconds off the actual travel time. The RMSE decreased to 144, indicating that model 3 was better at handling larger errors. The R^2 increased to 72.2%, which captures 9% more variation in travel time. These improvements confirm that the granularity of *timepoint segment* allowed the model to capture the nuances of the different segments, which led to more accurate predictions.

SHAP Feature Profiling

With Model 3 as the best performing model, we profiled SHAP features with the test dataset. In order to gain a better understanding of how the model produces a single prediction, SHAP values were employed to break down the predicted travel time into contributions for each feature. Figure 14 shows a SHAP waterfall plot for a sample prediction for one timepoint segment 540 during a single trip along the "36 Westbound WB" route. The baseline, in this case around 620 seconds, is the mean predicted *travel time* of the test dataset.

To calculate the final predicted *travel time* of 637 seconds for this representative sample prediction, each of the features adds positively or negatively to the baseline. The *traffic travel time* (582 seconds) is relatively high compared to the average of 387 seconds and therefore had the largest SHAP contribution of 145 seconds added to the predicted *travel time*. *Timepoint segment* 540 corresponds with the categorical code 14 in Figure 14. This timepoint has a shorter distance at 2.01 miles compared to the average of 3.03 miles for all timepoints in the data, thus we see a SHAP contribution of -82 seconds to the estimated *travel time*. *Passenger activity*, *length of service*, and *made stop ratio* also affected the *travel time*, though their SHAP values were smaller. There were 2 passengers, the vehicle served 1851 days, and the made stop ratio was .083, all of which were below their average values in the dataset (4, 2343, and .17 respectively). These values resulted in decreases to overall predicted travel time ~ 13 seconds from *passenger activity*, 13 seconds from *length of service*, and 10 seconds from *made stop ratio*.

These SHAP explanations provided at the individual prediction level facilitate an intuition regarding how individual features affect travel time predictions at the individual timepoint segment level. Through providing an interpretable picture of the extent to which different features affect travel time predictions, this example will help interpret how different features contribute to travel time predictions at the aggregate level.

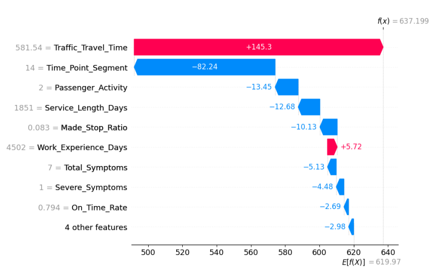


Figure 14: SHAP Waterfall Plot for Individual Prediction

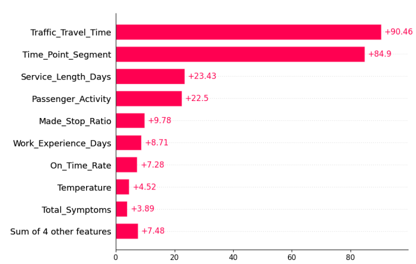


Figure 15: SHAP Summary Plot at Aggregate Level

The SHAP summary plot in Figure 15 illustrates the major factors affecting bus travel time predictions at the aggregate level. The aggregate contribution of each feature is measured from taking the average absolute value of the SHAP score, which measures the magnitude of the variable's effect regardless if it increases or decreases the predicted value. *Traffic travel time* is the most impactful variable because variation in traffic increases or decreases *travel time* estimates by 90 seconds on average. *Timepoint segment* follows in close second with a SHAP score of 85 seconds, meaning which timepoint the bus travels on can also have a significant effect on the predicted value. *Service length* and *passenger activity,* though not as impactful, contributed 3rd and 4th most to estimated *travel time* at approximately 23 seconds for both. This underscores that important features came from a wide variety of data sources.

Figure 16 displays a beeswarm plot for model 3 to provide insight into each feature's influence on the predicted *travel time*. *Traffic travel time* behaves as we would expect with lower values decreasing the predicted value and higher values increasing it. *Timepoint segment*, though impactful, is a cateogrial variable, so it is difficult to draw insights from this chart. *Length of Service* has mid-range values decrease the travel while higher range values increased it. Interestingly, low values had both positive and negative impacts, highlighting a limitation of SHAP that its interpreations are context dependent. For *length of service*, the same input value can have varying impacts based on other variables, which can be seen in the beeswarm plot. On the other hand, for *passenger activity* and *made stop ratio,* we see expected trends where higher values increase *travel time* and lower values decrease it. In contrast with model 1 where high *made stop ratios* were associated with shorter *travel times*, model 3 has high ratios being associated with longer *travel times*, which is more aligned with real-world dynamics.

Commented [KM15]: Says Figure 153

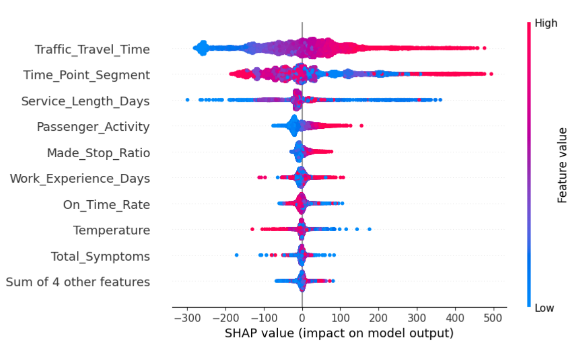


Figure 16: SHAP Beeswarm Plot for Model 3

Conclusion:

In order to create an XGBoost model that predicts segment-level bus travel time on MARTA's lowest performing routes, this study combined several internal and external data sources reflecting exogenous and exogenous factors influencing bus travel times. A process of early stopping and a robust hyperparameter tuning procedure were required to avoid overfitting, in which a model learns patterns that are too similar to the training data that it cannot apply them accurately to new data. Overfitting in this case would result in overly inaccurate travel time estimates, which would limit the applicability of the model to customer information systems and planning operations.

The performance of the model was excellent on a range of accuracy scores, and using SHAP gave useful insight into how each predictor affected individual and aggregate predictions. Transit agencies are greatly advantaged by this level of interpretability as it allows them to identify the most important drivers of delays. Ensuring that non-technical stakeholders can understand the decision-making process of the model, SHAP also promotes trust.

Even though the current model shows encouraging preliminary results, there are still limitations of this study. Only a subset of routes were analyzed; covering the entire network with the model will increase its applicability but requires a laborious manual mapping of RITIS segments to MARTA timepoints. Although the model currently forecasts travel time, on-time performance (OTP) is a metric that is more operationally relevant. Modeling runtime adherence, which is the discrepancy between actual and planned travel time, would be a sensible next step. Following the results from runtime adherence estimation, we can classify trips as early, late, or on-time. Future work can include additional features such as use of wheelchair ramp, fare payment methods, additional dwell periods, and whether bus dispatch communicated with an operator to assist or provide adherence-related feedback. These factors could enhance the model's predictive power and provide more detailed information about trip-level disruptions.

Commented [KM16]: Also seasonality.

Commented [KM17R16]: Time of day/ day of week if you haven't looked at this already (could you and let me know?)

References:

1. Zhu, L., Xu, J., Ding, M., & Li, C. (2022). XGBoost-based travel time prediction between bus stations and analysis of influencing factors. Wireless Communications and Mobile Computing, 2022, 1–25. https://doi.org/10.1155/2022/3504704
2. B. P., Ashwini., Ananthanarayanan, V., & Ramesh, B. (2022). A dynamic model for bus arrival time estimation based on spatial patterns using machine learning. International Journal of Engineering Trends and Technology, 70(9), 185–193. https://doi.org/10.14445/22315381/ijett-v70i9p219
3. Chiang, M.-L., Lin, C.-W., & Hsiao, W.-F. (2020). A survey of bus arrival time prediction methods. In Advances in Intelligent Systems and Computing (pp. 197–206). https://doi.org/10.1007/978-3-030-46828-6_17
4. Anggoro, D. A., & Mukti, S. S. (2021). Performance comparison of grid search and random search methods for hyperparameter tuning in extreme gradient boosting algorithm to predict chronic kidney failure. International Journal of Intelligent Engineering and Systems, 14(6), 198–207. https://doi.org/10.22266/ijies2021.1231.19
5. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5
6. Zhang, J., Wang, Y., Liu, S., & Chen, W. (2023). Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. Journal of Environmental Management, 332, 117357. https://doi.org/10.1016/j.jenvman.2023.117357

Appendix

A1: Timepoint and Timepoint Segment Definition

Timepoint: *A timepoint is a designated stop along a bus route where the bus is scheduled to depart at a specific time. All timepoints are bus stops, but not all stops are timepoints. Timepoints are the key locations used for schedule adherence and reporting, stops where departure times are explicitly defined in the published schedule. Stops between timepoints are served by the bus, but do not have scheduled times. Adherence data used to calculate OTP is always recorded for timepoints regardless of whether any passenger boards/alights the bus.*

Timepoint Segment: *A timepoint segment is the portion of a bus route between two consecutive timepoints. Travel time is typically measured across these segments to assess performance. Timepoint segments provide the unit of analysis for modeling travel times, as they align with MARTA's schedule structure and reporting framework.*

A2: TMDM to RITIS Ratio

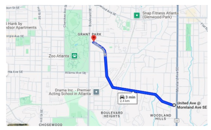
- RITIS Distance: The average length of the route segment covered by RITIS probe vehicles at the year, month, day, route, direction, and pattern level.
- TMDM Distance: the scheduled distance of a route segment, as defined by MARTA's TMDM system for this route-direction-pattern for a particular year-month-day.
- The TMDM to RITIS ratio indicates how completely the TMDM segment is represented within the RITIS data.

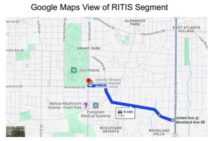
$$\text{TMDM to RITIS Ratio} = \frac{\text{TMDM Distance}}{\text{RITIS Distance}}$$

A3: RITIS Mapping Example

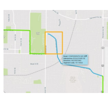
Traffic Behavior via RITIS Mapping

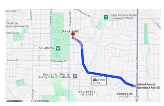
- MARTA Transit Analytics team previously reconstructed Bus Routes as RITIS segment sets to capture external traffic conditions.
- Each segment was mapped from RITIS segments to MARTA timepoint segments.
- Alignment between RITIS and TMDM segments were assessed to ensure accurate correspondence.
  - Applied multipliers (derived via Google Maps) to scale RITIS segment data where segment lengths differed from TMDM segment
  - Shorter RITIS segments > upscaled
  - Longer RITIS segments > downscaled
- This mapping allows us to generate traffic-based features at the timepoint segment level, enabling more precise modeling of how external traffic conditions impact MARTA bus travel time.

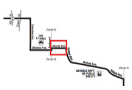

Google Maps View of RITIS Segment

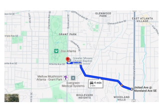

Google Maps View of TMDM Segment

RITIS Mapping Example


RITIS Segment Mapping for Route 832 Westbound

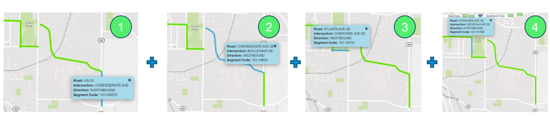

Google Maps View of RITIS Segment


Route 832 Westbound TMDM segment

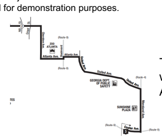

Google Maps View of TMDM Segment

- The RITIS segment extends beyond the route shown in TMDM.
  - Orange box highlights the extra portion of the TMDM segment.
- The RITIS segment is also missing part of the TMDM segment.
  - Red box highlights the missing portion in the TMDM segment.
- Both segments were mapped in Google Maps:
  - RITIS segment length: 2.4 km
  - TMDM segment length: 2.0 km
  - Ratio = 2.0 km / 2.4 km, therefore we apply a scaling factor of 5/6 to adjust the RITIS segment
- Assumption: Traffic conditions in the extra portion of the RITIS segment are assumed to be comparable to the traffic in the missing portion of the TMDM segment.
- The 5/6 scaling factor adjusts the RITIS segment's travel time to match the actual length of the corresponding TMDM segment.

RITIS Mapping Example

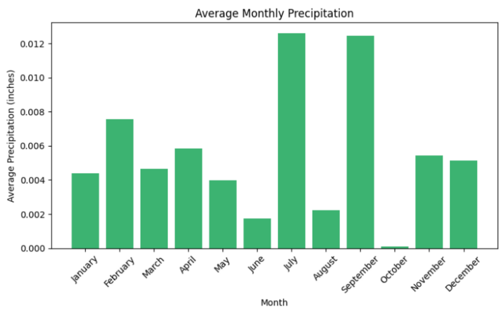


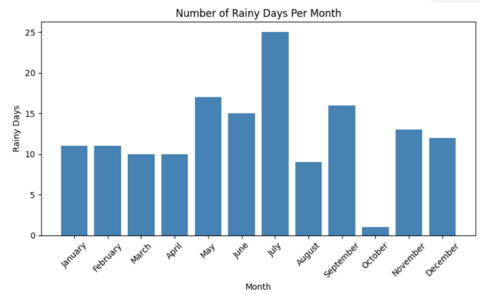
Using the process from the previous slide, the RITIS segments shown above are mapped to the TMDM timepoint segment below. For each timepoint segment, all corresponding RITIS segments are combined to calculate the total observed RITIS distance. Thi s allows us to compute a scaling factor that adjusts the combined RITIS traffic data to match the full TMDM timepoint segment. This example is simplified and only used for demonstration purposes.

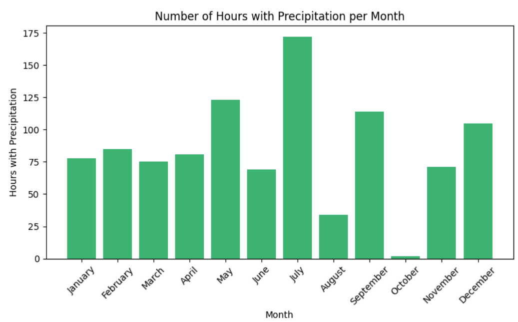


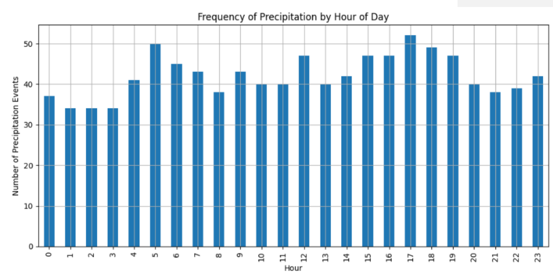
TMDM Timepoint Segment for route 832 westbound from Woodland Ave & Custer Ave (3) to Georgia Ave & Hill St (2)

A4: Precipitation Feature Characterization

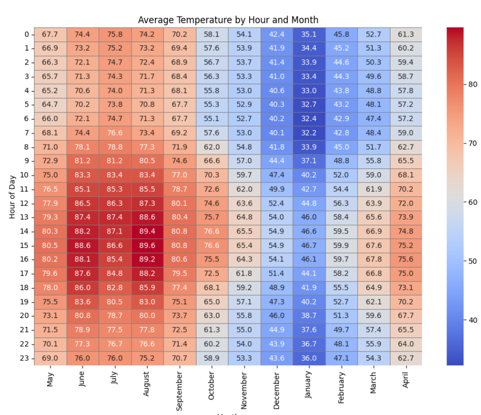




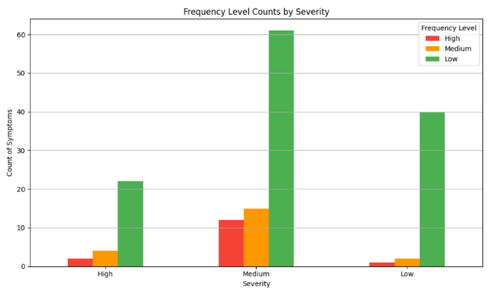


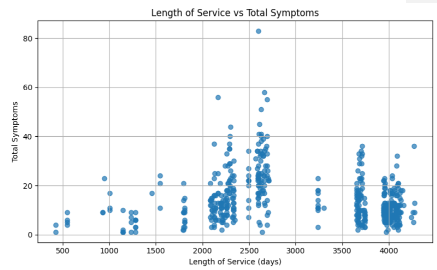


A5: Temperature Feature Characterization

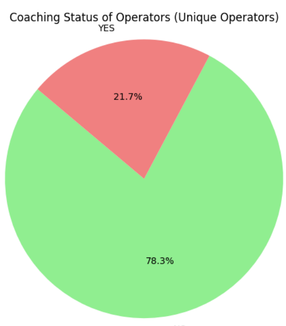


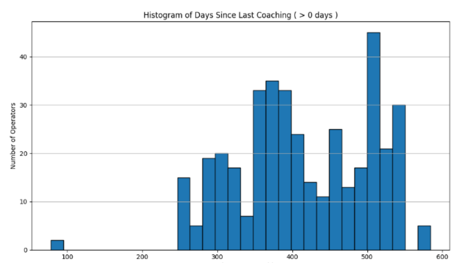
A6: Bus Maintenance Feature Characterization



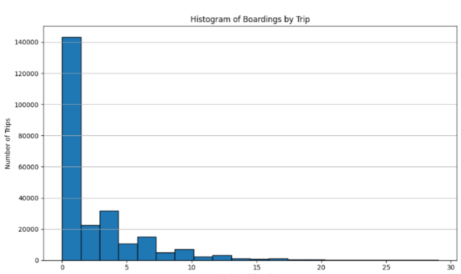


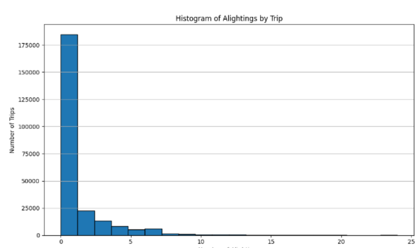
A7: Operator Performance Feature Characterization



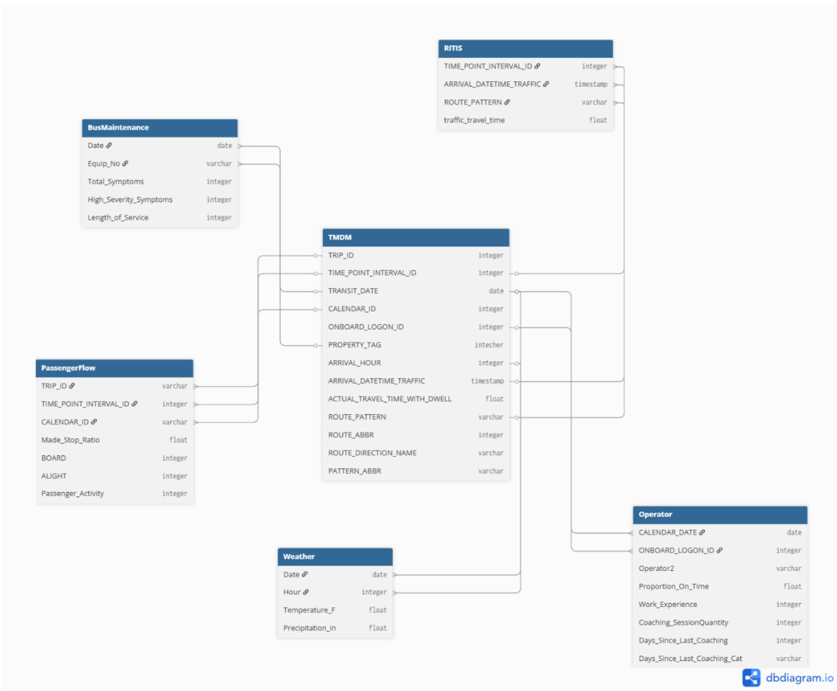


A8: Passenger Flow Feature Characterization





A9: Data Consolidation Flow Diagram



Not all fields are shown; only representative attributes are included for illustration purposes.

A10: Model 3 Residual Analysis